Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Ticino: A multi-modal remote sensing dataset for semantic segmentation

Mirko Paolo Barbato [a,*], Flavio Piccoli [a], Paolo Napoletano [a,b]

[a] *Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy*
[b] *Istituto Nazionale di Fisica Nucleare (INFN), Milano, 20126, Italy*

ARTICLE INFO

Keywords:
Multi-modal remote sensing
Image semantic segmentation
Deep learning
Multi-modal dataset
Data fusion
Hyperspectral imaging

ABSTRACT

Multi-modal remote sensing (RS) involves the fusion of data from multiple sensors, such as RGB, Multispectral, Hyperspectral, Light Detection and Ranging, Synthetic Aperture Radar, etc., each capturing unique information across different regions of the electromagnetic spectrum. The fusion of different modalities can provide complementary information, allowing for a comprehensive understanding of the Earth's surface.

Multi-modal RS image segmentation leverages various RS modalities to achieve pixel-level semantics classification. While deep learning has demonstrated promise in this domain, the limited availability of labeled multi-modal data poses a constraint on leveraging data-intensive techniques like deep learning to their full potential. To address this gap, we present Ticino, a novel multi-modal remote sensing dataset tailored for semantic segmentation.

Ticino includes five modalities, including RGB, Digital Terrain Model, Panchromatic, and Hyperspectral images within the visual-near and short-wave infrared spectrum. Specifically annotated for Land Cover and Soil Agricultural Use, the dataset serves as a valuable resource for researchers in the field. Additionally, we conduct a comparative analysis, comparing single-modality with multi-modality deep learning techniques and evaluating the effectiveness of early fusion versus middle fusion approaches.

This work aims to facilitate future research efforts in the domain by providing a robust benchmark dataset and insights into the effectiveness of various segmentation approaches.

1. Introduction

Remote Sensing (RS) is one of the most significant sources of information for the understanding of the land and its properties. Utilizing sensors mounted on drones, aircraft, or satellites, RS captures images of the Earth's surface from a distance. This process enables the monitoring and study of our planet's environment and its changes over time through advanced computer vision techniques.

The evolution of RS technologies has granted us access to diverse types of data, conveying rich and complementary information such as spectral data (Multispectral - MS and Hyperspectral - HS), Light Detection And Ranging (LiDAR), and Synthetic Aperture Radar (SAR). This variety underscores the complexity of Earth's components, facilitating improved investigations and resource management. For instance, spectral information is crucial for identifying specific materials, while terrain elevation data provides essential morphological insights. Although these data types individually offer new perspectives and capabilities, their full potential is realized when combined, thus optimizing the information obtained from each and compensating for any single source limitations.

In computer vision, the development of deep neural networks has not only demonstrated exceptional performance in tasks like classification, segmentation, and parameter regression but has also enabled the advancement of multi-modal approaches. These techniques optimally combine information from different modalities to extract the most valuable features for the task at hand.

In RS, semantic segmentation is a primary and essential analysis used in a wide array of applications, including autonomous driving, robot navigation, industrial inspection, saliency object detection, agricultural sciences, medical imaging analysis, and remote sensing itself (Lateef & Ruichek, 2019). This technique involves classifying each pixel in an image, resulting in a map that groups pixels into areas of the same semantic class (Yuan, Shi, & Gu, 2021). In RS, semantic segmentation plays a significant role in fields such as precision farming, environmental monitoring, spatial planning, and management of ecosystem-oriented natural resources (Blaschke, 2010; Dechesne, Mallet, Le Bris, & Gouet-Brunet, 2017; Jadhav & Singh, 2018; Kussul, Lavreniuk, Skakun, & Shelestov, 2017; Rottensteiner et al., 2012).

Despite the advancements in multi-modal approaches (Palhamkhani et al., 2023), their application in RS semantic segmentation is not yet

* Corresponding author.
E-mail addresses: mirko.barbato@unimib.it (M.P. Barbato), flavio.piccoli@unimib.it (F. Piccoli), paolo.napoletano@unimib.it (P. Napoletano).

https://doi.org/10.1016/j.eswa.2024.123600
Received 16 November 2023; Received in revised form 29 February 2024; Accepted 1 March 2024
Available online 5 March 2024

fully explored. RS semantic segmentation could greatly benefit from data types like hyperspectral images, which, despite lower spatial resolution, offer superior discrimination power due to their higher spectral resolution and band count (Barbato, Napoletano, Piccoli, & Schettini, 2022). However, the lack of multi-modal datasets, particularly those incorporating high spectral resolution data like hyperspectral images, is a significant challenge. This scarcity, mainly caused by difficulties in creating comprehensive semantic segmentation labelings and ensuring compatibility between different modalities, limits the full potential of multi-modal approaches and the technologies available to us. Nonetheless, the literature on RS data fusion has seen a significant increase in recent years, focusing on both homogeneous and heterogeneous fusion of complementary information, thus highlighting the importance of advancing research in this area (Li et al., 2022; Loncan et al., 2015; Vivone, Garzelli, Xu, Liao, & Chanussot, 2022).

In this research, we present the Ticino dataset, a novel satellite multi-modal remote sensing dataset specifically tailored for semantic segmentation tasks. This dataset fuses color, spatial, spectral, and morphological information across five modalities: RGB, Digital Terrain Model (DTM), Panchromatic, and Hyperspectral (HS) images. It spans the visual-near infrared and short wavelength infrared portions of the electromagnetic spectrum. With valuable spatial and color information from the RGB modality and effective material discrimination from the hyperspectral components, the dataset enhances our understanding of soil morphology. Notably, it includes labeled data for Land Cover and Soil Agricultural Use, covering an area of about 1332 km^2. To the best of our knowledge, it is the largest and most diverse multi-modal dataset for RS semantic segmentation.

We have conducted a comparative analysis to evaluate single-modality and multi-modality deep learning techniques, as well as early and middle fusion methodologies. Our findings demonstrate the superiority of multi-modal approaches, with middle fusion showing the most significant improvement in performance.

The main contributions and findings of this research related to remote sensing and semantic segmentation are:

- a multi-modal Remote Sensing dataset that combines RGB, Hyperspectral, and Digital Terrain Model, with both high spatial and high spectral resolutions;
- a baseline comparison of single- vs. multi-modality deep learning techniques, as well as early vs. middle data fusion techniques;
- empirical evidence that multi-modality enhances semantic segmentation accuracy compared to single modalities;
- empirical evidence that multi-modality is more effective when employing a middle fusion strategy;
- empirical evidence of the particular effectiveness of hyperspectral data in Soil Agricultural Use.

2. State of the art

In this section, we explore two critical aspects: the existing remote sensing (RS) datasets for semantic segmentation and the advancements in semantic segmentation methods within computer vision and RS. The first part highlights the lack of comprehensive RS multi-modal datasets, especially those incorporating hyperspectral (HS) data. The second part focuses on showing how even with a scarcity of multi-modal datasets properly built for segmentation, this field is still one of the most studied and challenging.

2.1. RS datasets for semantic segmentation

Despite the proliferation of RS data in terms of quantity, modality, and diversity, there is a notable disparity in dataset quality for semantic segmentation compared to other tasks, such as classification (Santiago, Schenkel, Gross, & Middelmann, 2020). The most significant datasets for RS semantic segmentation are categorized based on the type of data they encompass.

Three bands datasets. Focusing on RGB images, Deepglobe (Demir et al., 2018) is a notable dataset that covers an area of 1716.9 km^2, including Thailand, Indonesia, and India. This dataset includes labeling for land cover and land use segmentation. The TorontoCity dataset (Wang et al., 2016), which combines RGB and LiDAR information, is geared toward building footprints and road segmentation. It covers an area of 712.5 km^2. Other datasets like the SpaceNet variant (Mohanty et al., 2020) (only RGB components of the standard SpaceNet dataset (Van Etten, Lindenbaum, & Bacastow, 2018)) and the INRIA aerial dataset (Maggiori, Tarabalka, Charpiat, & Alliez, 2017) focus primarily on binary segmentation of buildings and non-buildings. Another RGB dataset is the Urban dataset from the Campinas region in Brazil (dos Santos et al., 2014). This dataset focuses on dividing urban and non-urban areas. The same article also presents the Coffee dataset (dos Santos et al., 2014) that considers images of 3 bands using the NIR-R-G part of the spectrum instead of the classical RGB. The dataset focuses on the detection of manually segmented coffee crops.

Multispectral datasets. These datasets incorporate multispectral modality either alone or in combination with other data types. They are crucial for extracting detailed information from the spectrum. The Zurich Summer dataset (Volpi & Ferrari, 2015), for example, includes four bands in the NIR-RGB part of the spectrum and a spatial resolution of 0.61 m per pixel achieved after the application of a pansharpening technique. The labeling represents eight urban classes.

Multi-modal datasets with multispectral information. SpaceNet (Van Etten et al., 2018; Yuan et al., 2021), which consists of images from different sensors including WorldView-1, WorldView-2, WorldView-3, WorldView-4, and GeoEye-1 (Arora, 2018), represents a key example of this category. Each sensor presents a variety of data and the most complementary between them is the WorldView-3 which includes a panchromatic image and 8 multispectral data, respectively in the VNIR and SWIR portions of the spectrum. It covers different cities and presents various kinds of segmentation depending on the type of task aimed.

Other multi-modal datasets are 2D Semantic Labeling Potsdam and Vaihingen datasets (Isp, 2023) that present both multispectral/RGB and DSM information with heights of the surface for each pixel. Potsdam also involves more versions of the same ground tiles. It includes two three-band images in the RGB or the IR-RG part of the spectrum and a third multispectral image of 4 bands with IR-RGB information that comprehends all the spectral information of the dataset. Vaihingen instead presents only RGB information when it comes to the spectrum. The two datasets present a labeling that includes six classes.

Another multi-source dataset that includes multispectral information is the DSTL dataset (Dst, 2023). It includes an RGB image from Deepglobe (Demir et al., 2018), a one-band panchromatic image, an eight-band multispectral image with NIR and visible information, and an eight-band multispectral image in the short wavelengths part of the spectrum. The dataset is built to identify 10 classes.

Hyperspectral datasets. Characterized by their high spectral resolution, hyperspectral datasets like Indian Pines (Baumgardner, Biehl, & Landgrebe, 2015), Salinas, SalinasA (M Graña & Veganzons, 2020), Pavia Center, and Pavia University (M Graña & Veganzons, 2020) are limited in data quantity and diversity, which affects their applicability for modern deep learning techniques. These datasets are primarily used for more specialized studies.

The Indian Pines dataset acquired by the AVIRIS sensor consists of 145 × 145 pixels and 224 spectral bands in the 400–2500 nm wavelength range. The final number of bands is reduced to 200 by removing the region of water absorption bands. The available ground truth includes sixteen classes, mainly regarding the distribution of different agricultural crops.

Salinas and SalinasA have been acquired by the AVIRIS sensor as well, and present 224 bands in the 400–2500 nm portion of the

Table 1
Comparison between state-of-the-art datasets for RS semantic segmentation and our Ticino dataset. Note that we present two versions of the dataset. One in the original scale and one at higher resolution obtained through the cleaning procedure and the pansharpening processing described in Section 3.2.

Dataset	Sensor	Modalities	Area [km^2]	# images	Image size	Res. [m/pixel]	# bands	# classes
Deepglobe (Demir et al., 2018)	Airborne	RGB	1716.9	1 156	2448 × 2448	0.50	3	7
TorontoCity (Wang et al., 2016)	Airborne	RGB/LIDAR	712.5	–	–	0.10	3/1	3
SpaceNet variant (Mohanty et al., 2020)	Satellite	RGB	3254[a]	401 755	300 × 300	0.3	3	2
INRIA (Maggiori et al., 2017)	Airborne	RGB	810	360	1500 × 1550	0.30	3	2
Urban dataset (dos Santos et al., 2014)	Airborne	RGB	3.46[a]	9	1000 × 1000	0.62	3	2
Coffee dataset (dos Santos et al., 2014)	Airborne	NIR-RG	56.25[a]	9	1000 × 1000	2.50	3	3
Zurich Summer (Volpi & Ferrari, 2015)	Satellite	MSI NIR-RGB	8.56	20	1000 × 1150	0.61	4	8
Indian Pines (M Graña & Veganzons, 2020)	Airborne	HSI VNIR-SWIR	0.29[a]	1	145 × 145	3.70	200	16
Salinas (M Graña & Veganzons, 2020)	Airborne	HSI VNIR-SWIR	1.52[a]	1	512 × 217	3.70	204	16
SalinasA (M Graña & Veganzons, 2020)	Airborne	HSI VNIR-SWIR	0.10[a]	1	86 × 83	3.70	204	6
Pavia Center (M Graña & Veganzons, 2020)	Airborne	HSI Visible	2.03[a]	1	1096 × 1096	1.30	102	9
Pavia University (M Graña & Veganzons, 2020)	Airborne	HSI Visible	0.63[a]	1	610 × 610	1.30	103	9
SpaceNet (Van Etten et al., 2018) (Arora, 2018; Yuan et al., 2021)	Satellite	PAN MSI VNIR MSI SWIR	3011	24 586	650 × 650	0.31 1.24 (orig.) 1.24 (orig.)	1 8 8	2
ISPRS Potsdam (Isp, 2023; Yuan et al., 2021)	Airborne	MSI IR-RGB PAN DSM	3.42[a]	38	6000 × 6000	0.05 0.05 0.05	4 1 1	6
ISPRS Vaihingen (Isp, 2023; Yuan et al., 2021)	Airborne	MSI IR-RGB PAN DSM	1.34[a]	33	2500 × 2000	0.09 0.09 0.09	4 1 1	6
DSTL (Dst, 2023)	Airborne	RGB PAN MSI VNIR MSI SWIR	1	57	–	0.50 0.31 1.24 7.50	3 1 8 8	10
Ticino/Our	**Satellite**	**RGB** **PAN** **HSI VNIR** **HSI SWIR** **DTM**	**1331.721**	**1 502**	**256 × 362** **96 × 192** **16 × 32 (96 × 192)** **16 × 32 (96 × 192)** **101 × 203**	**1.86-2.64** **5** **30 (5)** **30 (5)** **5**	**3** **1** **63 (60)** **171 (122)** **1**	**8/10**

[a] Computed using the other information in the table.

spectrum, like Indian Pines. The final datasets have 204 bands because the 20 noisy channels in the region of water absorption have been discarded. The two images have a size, respectively, of 512 × 217 and 86 × 83 pixels. In particular, the SalinasA dataset represents a subset of the Salinas dataset. Consequently, the labeling is different between the two datasets. Salinas is annotated with 16 classes representing the region of cultures, while SalinasA is annotated with six classes.

Pavia Center and Pavia University datasets have been acquired through the ROSIS sensor and have, respectively, a resolution of 1096 × 1096 and 610 × 610. Images present 102 and 103 channels. In both cases, a portion of the samples has been discarded because of missing information, resulting in two images, respectively, of 1096 × 715 and 610 × 340 pixels. The labeling includes nine classes for both datasets, representing land cover.

In the context of RS semantic segmentation, leveraging various modalities can enhance performance significantly (Li et al., 2022). However, the prevalent challenge of generating comprehensive semantic labelings (Barbato et al., 2022) often results in existing remote sensing datasets for semantic segmentation being limited to single modalities. Table 1 summarizes the characteristics of each dataset in more detail. These factors underscore the necessity for developing a dataset that fully harnesses the benefits of multi-modal approaches, particularly utilizing more discriminative data sources. The Ticino dataset introduced in this study has been specifically designed to overcome these challenges.

2.2. Deep learning multi-modal approaches for semantic segmentation

Multi-modal approaches are increasingly prevalent across various fields, including medical analysis, language translation, image annotation, and RS monitoring (Gao, Li, Chen, & Zhang, 2020; Jiang, Ma, Xiao, Shao, & Guo, 2021). These approaches involve combining different data sources to leverage the unique advantages of each type, enhancing the overall analysis. In the context of semantic segmentation, the fusion of modalities depends on the dataset, chosen model, and the fusion technique applied. Convolutional Neural Networks (CNNs) and, more recently, Transformer architectures are the predominant models in computer vision for analyzing images.

In remote sensing, fusion techniques exploit the benefits of diverse and complementary modalities. While deep learning, particularly Transformer architectures, has emerged as a significant choice (Li et al., 2022), their application in semantic segmentation of RS images remains relatively underexplored (Aleissaee et al., 2023). These fusions can be categorized into heterogeneous and homogeneous types (Li et al., 2022). Heterogeneous fusions involve combining modalities with different meanings, such as hyperspectral, LiDAR, and DTM. Homogeneous fusions, such as spatio-temporal fusion and pansharpening (Li et al., 2022; Loncan et al., 2015; Vivone et al., 2022), combine modalities of the same type. The latter is often used for upscaling multispectral images and recently for enhancing the spatial resolution of hyperspectral data to match the higher resolution of the panchromatic component.

Typically, with CNNs, the methods can be divided into 3 groups: early fusion (or data-level), middle fusion, and late fusion (Li, Zhang, Cheng, Huang, & Tan, 2017). The main difference lies in the stage of the CNN model where the modalities are concatenated. Early fusion combines sources at the beginning, late fusion merges high-level features from each modality independently at the end, and middle fusion represents an intermediary approach.

With the rise of Transformers, new architectures designed for fusion have been explored, leading to various strategies (Xu, Zhu, & Clifton, 2023). For instance, the fusion of the tokens, performed by summation or concatenation, can characterize the multi-modal approaches with Transformers (Gavrilyuk, Sanford, Javan, & Snoek, 2020; Parida, Srivastava, & Sharma, 2022). Following the concatenation approach, hierarchical attention represents another example of multi-modal techniques with transformers. It consists of concatenating and splitting the tokens before or after the attention mechanism.

The hierarchical attention can be categorized into two types based on the application order of the concatenation and splitting operations (from multi-stream to one-stream or vice versa) (Lin et al., 2020). Another strategy modifies the structure of the self-attention mechanism. One common approach among these strategies is the Cross Attention (Lu, Batra, Parikh, & Lee, 2019), which exchanges the query of one modality with another during the computation of the traditional attention mechanism. Finally, combinations of these techniques used concatenation and Cross Attention together (Hasan et al., 2021; Zhan et al., 2021).

In the realm of semantic segmentation for land use and agriculture applications, the incorporation of point cloud analysis methodologies, as proposed by Xie, Wang, Lu et al. (2023) and Xie, Wang, Wang et al. (2023), adds a valuable dimension to enhance the precision and contextual understanding of the spatial features involved in the segmentation process.

3. Materials and methods

In this section, we describe the data collected and the methods used to evaluate the advantages of multi-modal approaches in this context. This section is organized into two subsections. The first one describes the development and the characteristics of our novel multi-modal remote sensing dataset. The second one explains the general methodology used in our experiments to test the usefulness of each modality.

3.1. Data collection

The Ticino multi-modal satellite dataset comprises data collected from various sources, specifically:

1. RGB data from Microsoft Bing Maps (mic, 2023) (see Fig. 1(a));
2. panchromatic and hyperspectral data from ASI PRISMA (pri, 2023) (see Fig. 1(b), (c) and (d));
3. digital terrain model (DTM) of the area considered from Geoportal of Lombardia Region (geo, 2023) (see Fig. 1(e)).

The dataset also includes two different pixel-level labelings for semantic segmentation:

1. Land Cover collected from OpenStreetMaps (ope, 2023) and Italian Agenzie delle Entrate (age, 2023) (see Fig. 1(f));
2. Soil Agricultural Use collected from the Geoportal of Lombardia Region (geo, 2023) (see Fig. 1(g)).

The proposed dataset considers a territory around the Ticino river in the south of Milan and has an extension of 1332 km^2. This area has been chosen for its heterogeneity in terms of terrain composition and geomorphological variety. To support data-driven methods such as deep learning, we divided the original dataset in 1808 smaller tiles. Among them, 306 have been discarded as they presented a number of labeled pixels inferior to 1%. The final dataset is therefore composed of 1502 georeferenced tiles. Each tile consists of five data sources and two pixel-level labelings. Fig. 1(a–g) show the original images. Fig. 1(h–n) show two examples of tiles extracted from the dataset. Fig. 1(o–u) show the same tiles after a post-processing operation known in the state of the art as pansharpening (Zini, Barbato, Piccoli, & Napoletano, 2023), that is used for increasing the spatial resolution of the hyperspectral data with the auxilium of the panchromatic information. The dataset has been split into training, validation, and test in percentages of 70%, 15%, and 15%, resulting in 1051 images for training, 225 for validation, and 226 for testing.

RGB data. Fig. 1(a) shows the RGB data included in our dataset. It has been collected from the Microsoft Bing Map service (mic, 2023) through an open-source tool.[1] These images present a different horizontal and vertical resolution. Specifically, they have a spatial resolution of 1.86 m/px for the vertical dimension and 2.64 m/px for the horizontal one. The RGB source is the data with the highest spatial resolution in the dataset. Each RGB image tile has a dimension of about 256 × 362 pixels.

Panchromatic data. Fig. 1(b) shows the panchromatic (PAN) data collected from the ASI PRISMA satellite (pri, 2023). PAN is a grey-level image in the visible part of the spectrum (400–700 nm). It has the highest spatial resolution of the dataset, namely 5 m/px. The original PRISMA and RGB data from Microsoft Bing presented a problem of georeference disalignment that we solved using the approach described in Appendix A.1. The final PAN tiles have a resolution of about 96 × 192 pixels.

Hyperspectral data. Visual and Near-Infrared (VNIR) and Short-Wave Infrared (SWIR) cubes (Fig. 1(c) and (d)) present a resolution of 30 m/px (the lowest spatial resolution of the dataset) and a spectral resolution of less than 12 nm. This data has been collected from ASI PRISMA satellite (pri, 2023) with the level-2D pre-processing, which is the highest level distributed and solves most of the acquisition problems related to the atmosphere, co-registration, etc. The VNIR data includes the spectral information of the visible and near-infrared parts of the spectrum, from 400 to 1010 nm. The VNIR cubes present 63 bands out of the original 66, as three bands did not contain valuable information. The SWIR component of the dataset represents the information in the short wavelength infrared part of the spectrum, from 920 to 2500 nm, with a portion of the spectrum that overlaps the VNIR information. The SWIR cubes contain 173 bands, but even in this case, the last two have been discarded due to the absence of valuable information. For each sample in the dataset, the hyperspectral cubes are image tiles of around 16 × 32 pixels. The same alignment transformation applied on the PAN image has been applied to align the VNIR and SWIR data. Moreover, we distribute a second version of the dataset where the hyperspectral cubes have been enhanced to reach the same spatial resolution of the PAN image using a pansharpening algorithm detailed in Section 3.1.1. The resulting hyperspectral images are at a spatial resolution of about 96 × 192 pixels.

[1] https://github.com/dakshaau/map_tile_download.

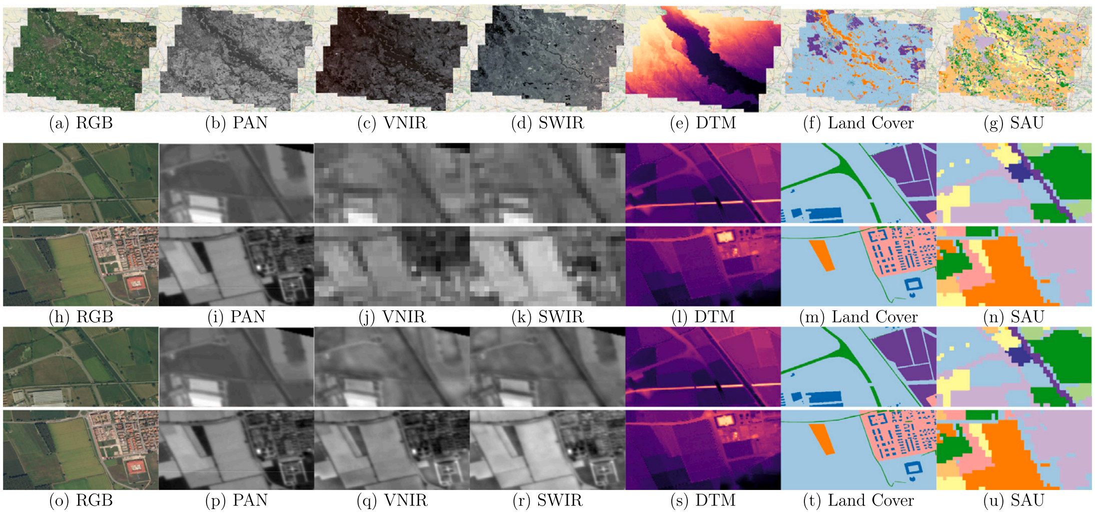


Fig. 1. Visual representations of each modality and labeling of the entire Ticino dataset (from (a) to (g)) and two examples of tiles (one tile for each row) with the corresponding multi-modalities (from (h) to (l)) and labelings ((m) and (n)).

Table 2
Land Cover classes of the presented Ticino dataset and image cardinality per class.

Classes	Id	# images
Background	0	1497
Building	1	1242
Road	2	1326
Residential	3	555
Industrial	4	216
Forest	5	675
Farmland	6	443
Water	7	169

Table 3
Soil Agricultural Use classes of the presented Ticino dataset and image cardinality per class.

Classes	Id	# images
Background	0	1475
Other agricultural crops	1	380
Forage crops	2	918
Corn	3	1029
Industrial plants	4	669
Rice	5	1323
Seeds	6	177
Man-made areas	7	1175
Water bodies	8	337
Natural vegetation	9	1315

Digital terrain model data. The last source included in our dataset is the Digital Terrain Model (DTM). As visible in Fig. 1(e), the DTM represents a topographic model of the bare Earth. It contains the elevation data of the terrain in a rectangular grid. It has been collected from the geoportal of the Lombardia region (geo, 2023). The DTM includes the urban and extra-urban areas. The model has been obtained from the geoportal by combining and harmonizing different sources of the data, removing possible anomalies, and finally extracting a Triangular Irregular Network model, achieving the final DTM model of the Lombardia region with a resolution of 5 m/px (geo, 2023). The DTM used in this dataset presents image tiles of about 101 × 203 pixels and an elevation that ranges from 51.86 to 124.75 m.

Land cover labeling. Fig. 1(f) shows the Land Cover segmentation. As with the RGB data, segmentation has different vertical and horizontal spatial resolutions of 0.68 and 0.96 m/px, respectively. The final labeling, which emerged from the refinement and merging process detailed in Appendix A.3, is a composite of data from several sources. These include Open Street Map (OSM) (ope, 2023), information from the Italian Agenzia delle Entrate (age, 2023), and additional labels that were manually annotated. At the end of the process, the dataset categorizes eight distinct classes: *Background*, *Building*, *Road*, *Residential*, *Industrial*, *Forest*, *Farmland*, and *Water*. The *Background* class represents unlabeled pixels. Table 2 shows the per-class cardinality in terms of number of images, along with the class name and identification number. The class distribution is slightly unbalanced, ranging from 169 images for the class *Water* to 1242 for the class *Building*.

Soil Agricultural Use labeling (SAU). Fig. 1(g) shows the SAU labeling. This segmentation has a resolution of 20 m/px (SIA, 2023) and it has been collected from the Geoportal of Lombardia region (geo, 2023). The labeling, after the refinements described in Appendix A.3, includes 10 classes: *Background*, *Other agricultural crops*, *Forage crops*, *Corn*, *Industrial plants*, *Rice*, *Seeds*, *Man-made areas*, *Water bodies*, and *Natural vegetation*. *Other agricultural crops* class indicates the not labeled farmlands and provides discrimination from the natural vegetation that instead describes forest, trees, and vegetation areas. Table 3 shows the image per-class cardinality of the Soil Agricultural Use, even in this case, along with the class name and identification number. As for Land Cover, the class distribution is slightly unbalanced ranging from 177 for the class *Seed* to 1323 for the class *Rice*.

3.1.1. Data pre-processing

As outlined in Section 3.1, we applied a pre-processing aimed at enhancing the quality of the hyperspectral (HS) component of our dataset. This process involved two key steps: the removal of corrupted bands and the enhancement of the spatial resolution of the HS data from 30 m/px to 5 m/px. Initially, we identified and discarded HS bands that contained no informative data. This preliminary screening resulted in a refined dataset comprising 63 bands for the Visible and Near-Infrared (VNIR) range and 171 bands for the Short-Wave Infrared (SWIR) range.

In line with the methodology proposed by Zini et al. (2023), the initial phase of the cleaning procedure targeted the corrupted bands of VNIR and SWIR within the HS component. This process used the information about invalid pixels described in the PRISMA documentation (He et al., 2023; pri, 2023). Each PRISMA image, in fact, comes with correspondent information regarding the validity of each pixel in each band. A pixel is not valid if a problem occurs during the acquisition phase or the PRISMA pre-processing. For each band, we compute the number of invalid pixels. Then, bands presenting a number of invalid

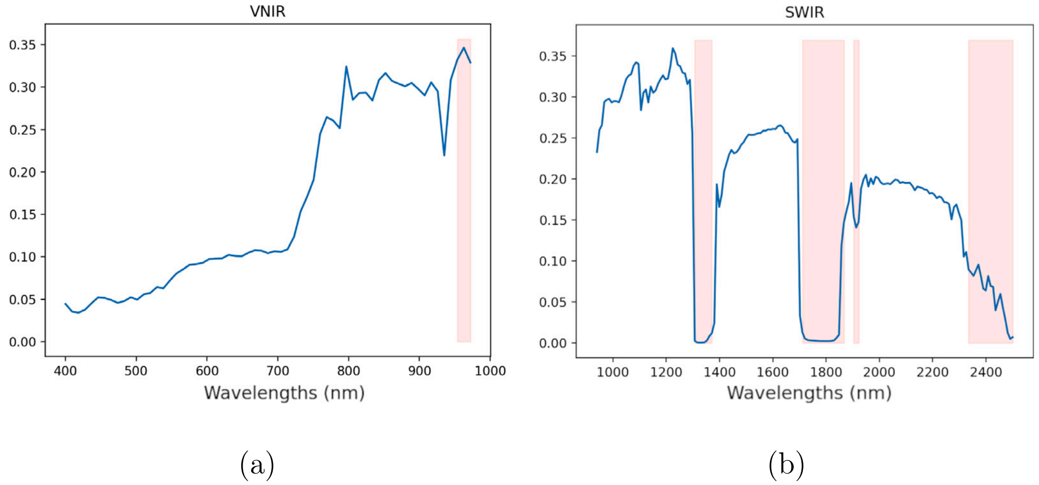


Fig. 2. Visual representation of the cleaning of the corrupted bands (first step of the pre-processing). The mean signature on each band with the removed bands from VNIR (left) and SWIR (right) pointed out in red. The figures show that the removed bands correspond to the overlapping band between the two modalities and the water absorption wavelengths (where the signal is almost zeroed out).

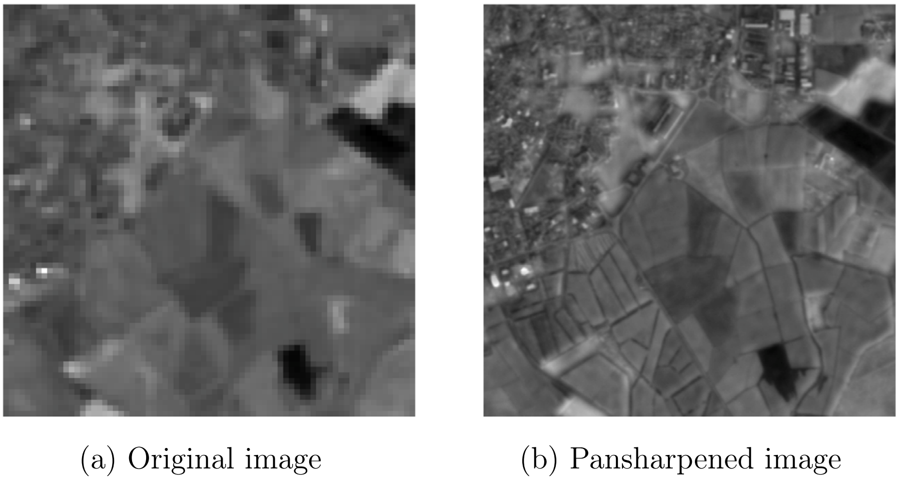


Fig. 3. Visual representation of the Panshaperning results (second step of the pre-processing). Comparison between the original hyperspectral image (band 50) on the left, and the hyperspectral pansharpened image (band 50) on the right after GSA algorithm.

pixels above a threshold, empirically fixed to 0.001%, are removed. The final part of VNIR is discarded, resulting in a new cube of 60 channels. The removed bands from SWIR mainly correspond to the water absorption part of the spectrum where the information is almost zeroed out.

In the second step, we conducted a visual inspection of each band. This inspection resulted in the removal of the 39th band of the SWIR component, due to the presence of visual artifacts. Consequently, the refined SWIR cube comprises 122 channels. To better visualize the effect of the cleaning procedure, Fig. 2 highlights in red the corrupted bands that were removed from the VNIR (Fig. 2(a)) and SWIR (Fig. 2(b)) signals.

To take advantage of PRISMA data, a pansharpening operation has been used to improve the spatial resolution of VNIR and SWIR. PRISMA satellite provides a panchromatic image (PAN) and two hyperspectral cubes for VNIR and SWIR information captured at the same time. Following the results of Loncan et al. (2015) and Vivone et al. (2022) on hyperspectral and PRISMA data pansharpening, the Gram–Schmidt Adaptive (GSA) (Aiazzi, Baroni, & Selva, 2007) algorithm has been selected. The pansharpening has been applied on VNIR and SWIR concatenated in a single hyperspectral data.

The final result (HS↑) is a hyperspectral cube corresponding to the fusion of the spectral information (VNIR and SWIR) and the spatial information from the PAN data. Fig. 3 shows the 50th band of the hyperspectral signal in its original form (Fig. 3(a)) and the same band after the pansharpening operation through the GSA algorithm (Fig. 3(b)). The output has a spatial resolution of 5 m/px (same as PAN) and a total of 182 bands that correspond to the VNIR and SWIR channels concatenated after the cleaning phase.

The final version of the dataset used in our experiments consisted of these modalities obtained by fusing PAN with VNIR and SWIR:

- RGB with 3 bands and a resolution of 1.86/2.64 m per pixel;
- Hyperspectral with 182 bands and a resolution of 5 m per pixel (HS↑);
- Digital Terrain Model (DTM) with 1 band and a resolution of 5 m per pixel.

3.2. Methods

In our experiments, we have considered different combinations and techniques of fusion for our modalities.

For each configuration, we have tested the same neural network model, consisting of a U-shaped architecture using the Segmentation Models PyTorch framework.[2] Our approach involves an encoder–decoder network with skip connections (the U-shaped architecture), where the ResNet18 serves as the encoder (detailly described in Appendix B). This choice was made intentionally, as ResNet18 remains a state-of-the-art architecture in visual recognition. By integrating ResNet18 into our network, we leverage its proven ability to capture complex hierarchical features, facilitating effective information extraction. Moreover, we have extended and customized the architecture with additional components, moving beyond a simple implementation. Specifically, the use of early fusion and middle fusion techniques is a deliberate design choice aimed at enhancing semantic segmentation performance.

The decision to use early or middle fusion in specific scenarios, in fact, is rooted in a careful consideration of the trade-offs between feature abstraction and spatial information preservation. Early fusion involves combining multi-scale features at the earliest layers of the network, allowing for a comprehensive integration of both low-level and high-level information. This is particularly beneficial when spatial details are critical for accurate segmentation. On the other hand, middle fusion occurs at intermediate layers, facilitating a balance between abstraction and detailed spatial information. This is advantageous in scenarios where capturing context and global information is essential for accurate segmentation tasks.

For every test, the same settings of the learning rate, data augmentation, and normalization have been considered using the Albumentations library (Buslaev et al., 2020). To train the models we have used 400 epochs, Adam optimizer with a learning rate of 0.0001, and a learning rate scheduler with *step size* of 30 and a decay gamma of 0.85, which means that the learning rate is updated and reduced every 30 epochs with a value equal the current learning multiplied by 0.85. The use of data augmentation, dropout, and early stopping have been used to reduce the problem of overfitting. In particular, the data augmentation for the training consisted of RGB normalization, HS↑ normalization, DTM normalization, a resize of each input to 256 × 352, a random crop of 256 × 256, random rotation applied between −180° and 180°, a random horizontal and vertical flip of the image, and finally a transpose transformation. The validation and test data augmentation consider only RGB normalization, HS↑ normalization, DTM normalization, and image resize to 256 × 352.

Every modality has been normalized using the z-score, that is removing the average and dividing by the standard deviation (Buslaev et al., 2020).

Fig. 4 shows the complete procedures for both early and middle fusion. Both start with the pre-process to clean hyperspectral data from corrupted bands and improve their spatial resolution using the Gram–Schmidt Adaptive (GSA) pansharpening technique described in 3.1.1.

Early fusion. As shown in Fig. 4(a), the pipeline for data-level fusion experiments consists of naively concatenating all the modalities together before using them as input of the U-shaped model.

The different combinations of modalities described above have 3 bands for RGB, 182 for HS↑, 185 for (RGB + HS↑), and 186 for (RGB + HS↑ + DTM). The definition of the U-shaped architecture and the layers of ResNet18 remained the same for all the experiments apart from the input layer which is changed according to the dimension of the input.

Middle fusion. In the middle fusion approach, as shown in Fig. 4(b), the different modalities are firstly processed independently to extract high-level features from each of them and later concatenated the features in order to create the input for the U-shaped architecture.

For each modality, the feature extraction module consists of convolutional and ReLU layers that use padding to maintain the same width and height of the U-shaped architecture. In the RGB and DTM cases, 3 convolutional layers increase the number of channels and extract the features. In the hyperspectral case, 2 convolutional layers are not only used to extract features but also to optimally reduce the number of channels of the starting hyperspectral inputs, thus overcoming the problem of the curse of dimensionality (Barbato et al., 2022). Table 4 summarizes the middle fusion module used for the extraction of the features in all modalities. After the application of the convolutional layers, all of the modalities share the same amount of feature maps to balance their importance during the training and are concatenated to become the input for the U-shaped architecture.

Table 4
Middle fusion module for the extraction of features from each modality.

Modality	Layer	Description	Padding
RGB	Conv2d	3 × 3 × 16	2 × 2
	ReLU		
	Conv2d	3 × 3 × 32	2 × 2
	ReLU		
	Conv2d	3 × 3 × 64	2 × 2
	ReLU		
Hyperspectral (HS↑)	Conv2d	3 × 3 × 128	2 × 2
	ReLU		
	Conv2d	3 × 3 × 64	2 × 2
	ReLU		
DTM	Conv2d	3 × 3 × 16	2 × 2
	ReLU		
	Conv2d	3 × 3 × 32	2 × 2
	ReLU		
	Conv2d	3 × 3 × 64	2 × 2
	ReLU		

4. Experiments and results

The configurations we have compared are:

- single-modality: RGB, HS↑ or DTM;
- multi-modalities: (RGB + HS↑), (RGB + DTM) or (RGB + HS↑ + DTM).

The two fusion techniques we have tested for each of the possible combinations are:

- early fusion;
- middle fusion.

The evaluation and comparison of the experiments are based on Accuracy (Acc), mean Intersection over Union (mIoU), and Precision. All of them have been computed by considering the average performance of single classes. A comprehensive evaluation of single classes is also reported, always based on the same metrics.

We have evaluated all modalities and fusion strategies on the two proposed labelings:

- Land Cover;
- Soil Agricultural Use.

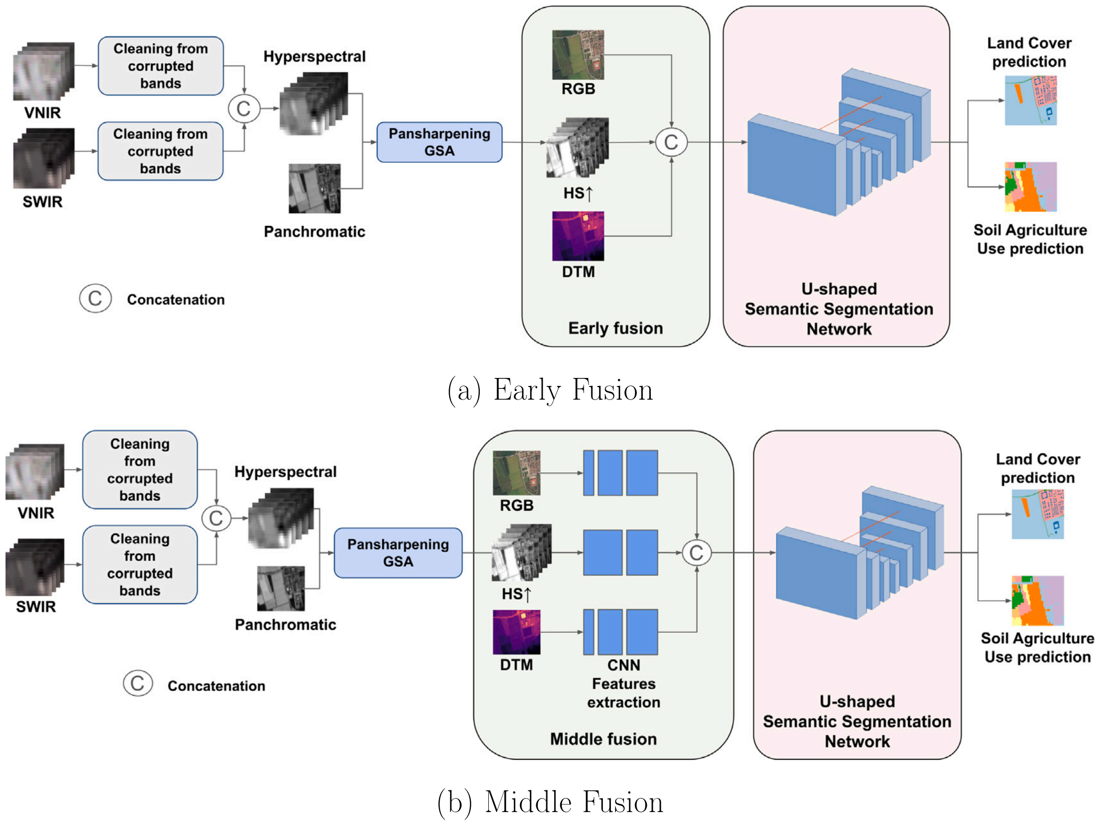
In the following, we present the results obtained in all the experiments as well as a discussion highlighting key findings that can be considered insights for future research in this area.

4.1. Land cover evaluation

Table 5 shows overall and class-specific results for Land Cover estimation achieved using different configurations.

Regarding single-mode experiments, RGB is the modality with the best overall performance in terms of accuracy, IoU and Precision. On the one hand, this is because the HS cube has a lower resolution than

[2] https://github.com/qubvel/segmentation_models.pytorch.



(a) Early Fusion

(b) Middle Fusion

Fig. 4. Experiments pipelines that represent the pre-processing, the fusion, and the segmentation. The two images show (a) the early and (b) the middle fusion techniques. In the early fusion technique, the data are concatenated immediately after the pre-processing and before the U-shaped network. In the middle fusion technique, the data are concatenated only after extracting high-level features using three ad-hoc CNNs, one for each modality independently. The concatenated features are then fed into the U-shaped network.

Table 5
Land Cover overall and single classes results of every experiment configuration divided by modalities combination and fusion techniques. Bold values represent the best performance obtained on the rows.

Land Cover										
Class	Metric	No fusion			Early fusion			Middle fusion		
		RGB	HS↑	DTM	RGB HS↑	RGB DTM	RGB HS↑ DTM	RGB HS↑	RGB DTM	RGB HS↑ DTM
Building	Acc	0.62	0.39	0.00	0.64	0.62	0.63	**0.75**	0.74	0.69
	IoU	0.50	0.31	0.00	0.49	0.49	0.48	**0.54**	**0.54**	0.53
	Prec.	**0.71**	0.60	0.36	0.68	0.70	0.67	0.66	0.68	0.69
Road	Acc	0.52	0.29	0.03	0.45	0.45	0.41	0.57	**0.61**	0.55
	IoU	0.42	0.23	0.03	0.34	0.38	0.33	0.41	**0.45**	0.41
	Prec.	**0.69**	0.52	0.23	0.58	**0.69**	0.62	0.59	0.64	0.61
Residential	Acc	0.85	**0.87**	0.50	0.82	0.85	0.85	0.75	0.76	0.80
	IoU	**0.64**	0.57	0.22	0.62	0.60	0.58	0.63	0.63	**0.64**
	Prec.	0.72	0.62	0.28	0.72	0.67	0.64	0.79	**0.80**	0.76
Industrial	Acc	0.64	0.52	0.50	0.62	0.64	0.47	**0.72**	0.69	0.67
	IoU	0.50	0.40	0.00	0.47	0.51	0.41	**0.55**	0.54	0.54
	Prec.	0.70	0.64	0.00	0.65	0.71	**0.75**	0.70	0.71	0.74
Forest	Acc	0.92	0.90	0.66	0.92	0.88	0.89	0.95	0.95	**0.96**
	IoU	0.87	0.85	0.51	0.88	0.83	0.86	0.90	0.90	**0.92**
	Prec.	0.94	0.93	0.69	0.95	0.94	**0.96**	0.95	0.95	**0.96**
Farmland	Acc	0.93	0.91	0.63	0.93	0.94	**0.95**	0.93	**0.95**	**0.95**
	IoU	0.85	0.82	0.39	0.86	0.83	0.88	0.87	0.89	**0.90**
	Prec.	0.91	0.89	0.51	0.91	0.87	0.92	0.94	0.93	**0.95**
Water	Acc	0.79	0.86	0.02	0.87	0.75	0.85	**0.89**	0.76	0.88
	IoU	0.65	0.72	0.02	**0.74**	0.66	0.73	**0.74**	0.65	0.73
	Prec.	0.79	0.82	0.06	0.83	0.83	**0.85**	0.81	0.82	0.81
Overall	Acc	0.75	0.68	0.26	0.75	0.73	0.72	**0.79**	0.78	0.78
	IoU	0.63	0.56	0.17	0.63	0.61	0.61	0.66	0.66	**0.67**
	Prec.	0.78	0.72	0.30	0.76	0.78	0.77	0.78	**0.79**	**0.79**

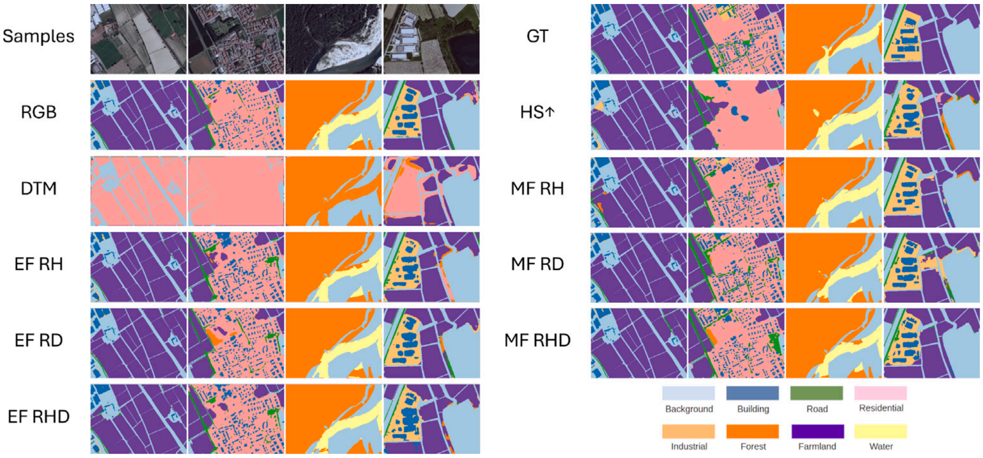


Fig. 5. Visual prediction of Land Cover segmentation for all the approaches. EF and MF are respectively for Early and Middle Fusion, while RH, RD and RHD are respectively for the combinations (RGB + HS↑), (RGB + DTM) and (RGB + HS↑ + DTM).

the RGB images, thus causing a loss of finer details in the segmentation process. On the other hand, the DTM does not carry enough information to allow a reliable estimation.

Concerning multi-modal experiments, the utilization of early fusion yields comparable results with the RGB modality. However, the middle fusion strategy is able to outperform it, suggesting that multi-modality in RS semantic segmentation is crucial. In particular, all middle fusion setups (RGB + HS↑), (RGB + DTM) and (RGB + HS↑ + DTM), which are very similar in terms of performance, outperform RGB by about 4%, 4%, 1% in terms of Accuracy, mIoU, and Precision respectively. Middle fusion configurations also outperform HS by about 11%, 11%, and 5% in terms of Accuracy, mIoU, and Precision, respectively.

The prevalence of middle fusion over early fusion can be attributed to the shallow nature of the latter, which hinders the optimal utilization of the distinctive attributes of each source.

As expected, the results in terms of mIoU on the single classes exhibit the same behavior of the overall performance. The classes which scored the lowest and the highest performance are respectively *Road*, with an IoU of 0.45 and *Forest*, with an IoU of 0.92.

It is worth noting that the best result of each class is obtained by combining different sources. While, for example, the *Road* class has a higher score when using (RGB + DTM), the Industrial class stands out more when using (RGB + HS↑). This behavior indicates that each semantic class benefits more from one source than another.

With the exception of the *Road*, *Farmland* and *Water* classes, where the modalities used by both fusion strategies agree, in all other classes the selected modalities are not concordant.

Fig. 5 shows visual results for all the considered approaches. Focusing on the best overall model (middle fusion with all modalities), it accurately classifies all labels, from fine-grained *Road* and *Building* to coarse-grained *Residential*, *Farmland*, and *Industrial* classes. Notably, the model recognizes a forest area located in the second RGB image but not in the correspondent labeling, demonstrating good performance even with noisy labels.

4.2. Soil agricultural use evaluation

Table 6 presents the overall results for Soil Agricultural Use, highlighting the poor performance of the single-modality RGB with respect to multi-modal approaches and the beneficial impact of HS modality in class discrimination. Due to the lower resolution of SAU labeling, we focus on evaluating the Accuracy rather than the mIoU. The early fusion approach already demonstrated the advantages of using multi-modality w.r.t. RGB (with an increment of 10%) and to HS (3%) modalities. DTM also contributes to segmentation, yielding improvements in accuracy and mIoU. As observed for Land Cover, the choice of fusion methodology is crucial. Middle fusion approaches showcase the true advantages of a multi-modal approach, outperforming single-modality experiments with the best results obtained by combining all modalities.

The difference in performance between middle fusion (RGB + HS↑ + DTM) and RGB is significant, with an increment of about 13%, 10%, and 13% for Accuracy, mIoU, and Precision, respectively. Similarly, the same middle fusion strategy scored a positive difference in performance with HS of about 6%, 4%, and 5% for Accuracy, mIoU, and Precision, respectively.

The improvement gained by using middle fusion with HS and DTM w.r.t. HS-only modality is about 6%, 4%, and 5% for Accuracy, mIoU, and Precision, respectively.

Table 6 also reports the segmentation results for each class. All multi-modal methods outperform RGB in class discrimination, once again confirming the importance of multi-modal approaches in RS semantic segmentation.

The *Seeds* class, in particular, demonstrates significant improvements when other modalities are utilized, going from 1% accuracy with RGB-only to 25% accuracy with all modalities and the middle fusion approach. Fusion methodology also plays a crucial role, with middle fusion generally yielding better improvements over early fusion.

Visual results for each combination and fusion technique are reported in Fig. 6. The segmentations achieved by the best approach, with middle fusion and all the modalities involved, accurately identify all classes despite the low resolution of SAU labeling.

This investigation demonstrates the usefulness of a multi-modal approach, especially for Soil Agricultural Use segmentation. Hyperspectral data and the Digital Terrain Model prove to be even more beneficial in this context than in Land Cover labeling, where RGB alone fails in achieving satisfactory results. Consequently, the availability of comprehensive multi-modal datasets is crucial for future research.

4.3. Discussion

Overall, experiments demonstrate the usefulness of multi-modality in semantic segmentation only when modalities are suitably combined with middle fusion strategies. Moreover, multi-modality is much more effective in the Soil Agriculture Use labeling rather than Land Cover, thus suggesting that hyperspectral and DTM modalities are much more effective on non-man-made classes. To visualize these findings, in Fig. 7 we show the percentage of increment/decrement in terms of the average of accuracy, IoU and precision when: (1) Multi vs Single - where we assess multi-modal approaches (regardless of the modalities used) against single-modality setups (irrespective of the chosen modality and fusion strategy); (2) Middle vs Early - where we compare the efficacy

Table 6
Soil Agricultural Use overall and single classes results of every experiment configuration divided by modalities combination and fusion techniques. Bold values represent the best performance obtained on the rows.

Soil Agricultural Use										
Class	Metric	No fusion			Early fusion			Middle fusion		
		RGB	HS↑	DTM	RGB HS↑	RGB DTM	RGB HS↑ DTM	RGB HS↑	RGB DTM	RGB HS↑ DTM
Other agricultural crops	Acc	0.26	0.31	0.00	0.34	0.31	0.36	0.42	0.45	**0.46**
	IoU	0.17	0.26	0.00	0.27	0.23	0.26	0.29	0.29	**0.32**
	Precision	0.34	0.63	0.00	0.55	0.49	0.48	0.50	0.46	**0.52**
Forage crops	Acc	0.14	0.32	0.00	0.32	0.21	**0.40**	0.35	0.22	0.37
	IoU	0.11	0.24	0.00	0.23	0.14	0.26	0.24	0.15	**0.27**
	Precision	0.34	**0.52**	0.00	0.45	0.33	0.42	0.44	0.32	0.48
Corn	Acc	**0.51**	0.45	0.01	0.48	0.36	0.47	0.49	0.48	**0.51**
	IoU	0.31	0.31	0.00	0.32	0.24	0.33	0.34	0.29	**0.36**
	Precision	0.45	0.50	0.11	0.49	0.43	**0.54**	0.52	0.42	**0.54**
Industrial plants	Acc	0.17	0.31	0.00	0.34	0.23	0.31	**0.46**	0.19	0.38
	IoU	0.09	0.18	0.00	0.19	0.13	0.19	**0.27**	0.12	0.21
	Precision	0.16	0.29	0.00	0.30	0.22	0.33	**0.39**	0.56	0.33
Rice	Acc	0.74	**0.81**	0.76	0.78	0.77	0.80	**0.81**	0.73	0.80
	IoU	0.57	0.64	0.42	0.63	0.58	0.65	**0.68**	0.57	**0.68**
	Precision	0.72	0.75	0.49	0.77	0.70	0.77	**0.81**	0.72	**0.81**
Seeds	Acc	0.01	0.06	0.00	0.12	0.08	0.17	0.21	0.23	**0.25**
	IoU	0.00	0.04	0.00	0.07	0.05	0.10	0.16	0.11	**0.20**
	Precision	0.02	0.14	0.00	0.12	0.12	0.20	0.39	0.17	**0.53**
Man-made areas	Acc	0.89	0.89	0.49	0.89	0.88	0.89	**0.90**	0.89	**0.90**
	IoU	0.77	0.76	0.28	**0.78**	0.76	0.76	0.77	0.76	0.77
	Precision	0.85	0.83	0.40	**0.86**	0.85	0.83	0.84	0.85	0.84
Water bodies	Acc	0.56	0.72	0.01	0.70	0.65	**0.75**	0.66	0.63	0.69
	IoU	0.46	0.55	0.00	0.56	0.52	**0.57**	0.55	0.51	0.56
	Precision	0.72	0.69	0.06	0.74	0.72	0.70	**0.77**	0.74	0.75
Natural vegetation	Acc	0.82	0.83	0.61	0.83	0.82	0.80	0.84	0.78	**0.85**
	IoU	0.64	**0.67**	0.36	**0.67**	0.65	0.65	**0.67**	0.65	**0.67**
	Precision	0.75	0.78	0.47	0.77	0.76	0.78	0.77	**0.79**	0.76
Overall	Acc	0.47	0.54	0.22	0.55	0.49	0.57	0.59	0.53	**0.60**
	IoU	0.35	0.41	0.12	0.41	0.37	0.42	0.44	0.38	**0.45**
	Precision	0.50	0.59	0.17	0.58	0.53	0.58	0.62	0.54	**0.63**

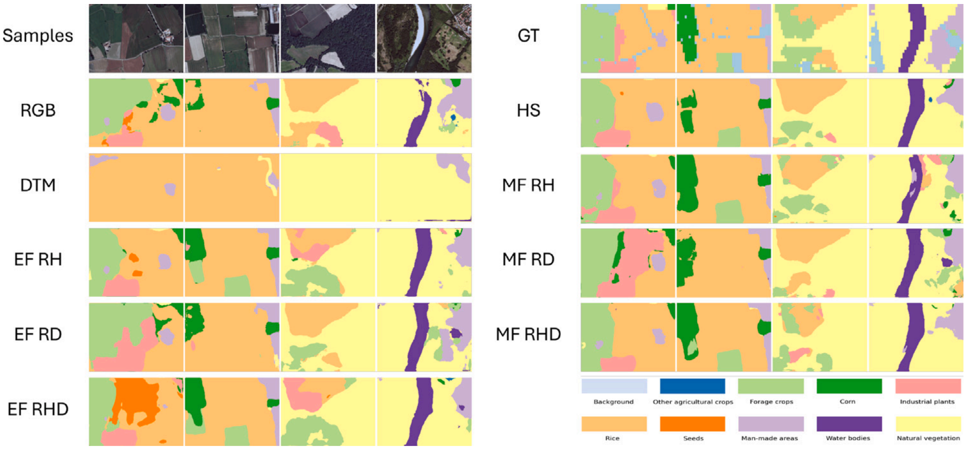


Fig. 6. Visual prediction of Soil Agricultural Use segmentation for all the approaches. EF and MF are respectively for Early and Middle Fusion, while RH, RD and RHD are respectively for the combinations (RGB + HS↑), (RGB + DTM) and (RGB + HS↑ + DTM).

of middle fusion (using any combination of modalities) against early fusion (with any modality combination).

In particular, figure (a) depicts the percentage of increment/ decrement in the case of Land Cover (LC), while figure (b) in the case of Soil Agriculture Use (SAU). In the case of LC, the average increment of using multi-modalities instead of single modalities is about 6%, and the average increment of using middle fusion instead of early modalities is about 6%. In the case of SAU, the average increment of using multi-modalities instead of single modalities is about 19%, and the average increment of using middle fusion instead of early modalities is about 10%.

To further highlight the advantages of multi-modality, it is also necessary to underline, as reported in Table 7, that the architectures used in the single-modality and multi-modality experiments share a similar number of parameters (around 15 Million parameters), thus the complexity required by these models does not suffer from the multi-modal approaches. Considering the GFLOPs, we can observe a

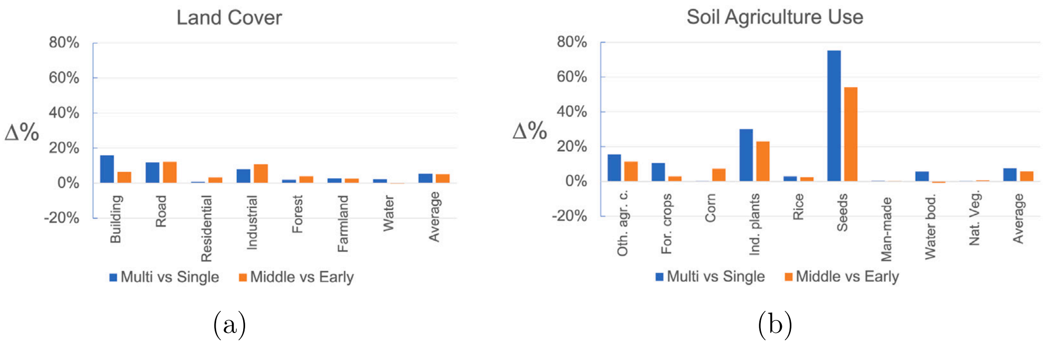
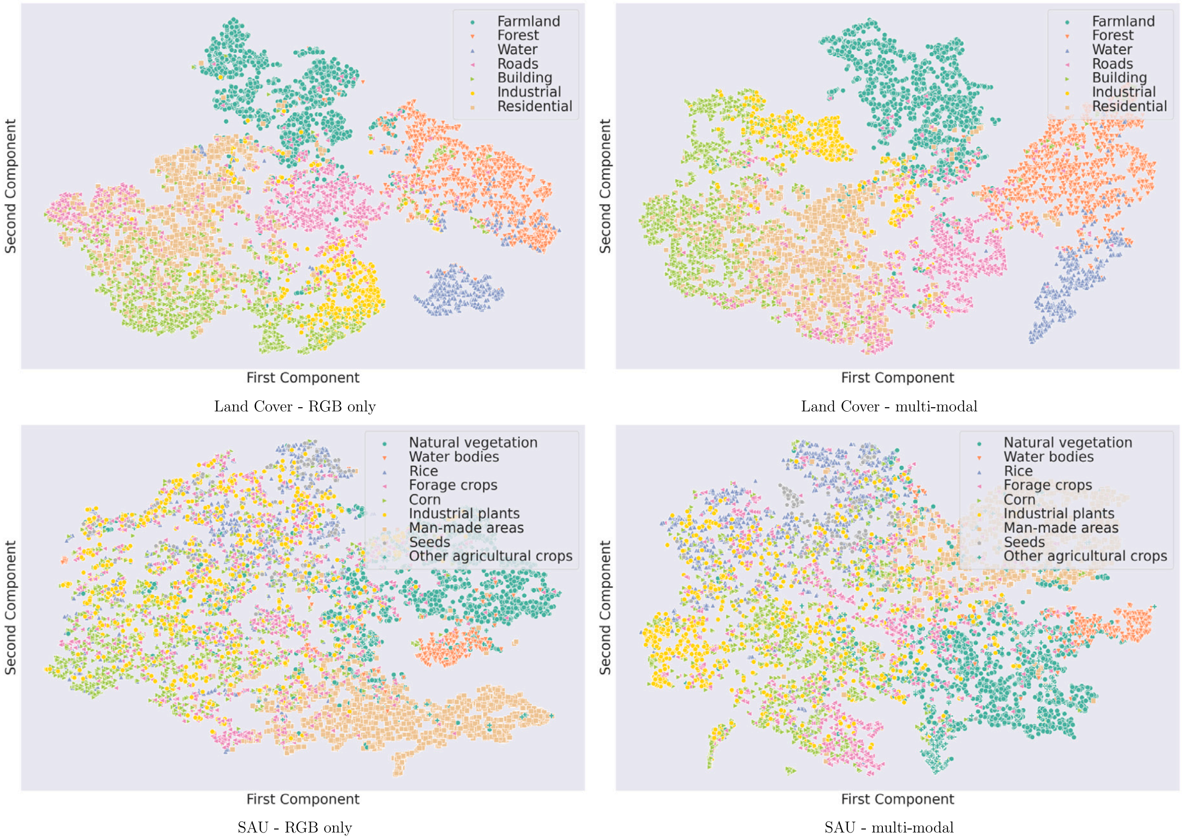


Fig. 7. Percentage of increment/decrement in terms of the average of accuracy, IoU and precision when: (1) we use multi-modal (whatever is the modalities adopted) with respect to single-modality (whatever is the single-modality combined and whatever is the fusion strategy); (2) we use middle fusion (whatever are the modalities combined) with respect to early fusion (whatever are the modalities combined). Figure (a) depicts the percentage of increment/decrement in the case of Land Cover, while Figure (b) in the case of Soil Agriculture Use.



Land Cover - RGB only

Land Cover - multi-modal

SAU - RGB only

SAU - multi-modal

Fig. 8. The first and second rows depict the TSNE representation of the features related respectively to Land Cover and Soil Agricultural Use. The first column contains features that are relevant to the RGB-only, while the second column represents the multi-modal features. It is evident that in both tasks, the utilization of multi-modality produces more representative features.

substantial increment in the number of operations required when the HS modality is employed, disregarding the fusion technique adopted.

The examination of the feature space further validates these conclusions. The analysis is visually represented in Fig. 8. The first and second rows illustrate the t-distributed stochastic neighbor embedding (t-SNE) representation of the features associated with Land Cover and Soil Agricultural Use, respectively. Both challenges demonstrate that the RGB-only features (first column) are less distinct compared to the multi-modal features (second column). This result reaffirms that employing multi-modality enhances the discerning capabilities of deep-learning-based systems in segmentation tasks.

5. Conclusions

In this paper, we have presented the Ticino dataset, a novel multi-modal dataset for RS semantic segmentation, that is crucial in various applications, including environment management and precision farming. The use of multi-modal sources of information enhances the segmentation performance and class discrimination, thus the scarcity of existing multi-modal datasets poses challenges in RS semantic segmentation. Existing datasets have low cardinality or lack spectral information, limiting the effectiveness of data-hungry deep-learning techniques that require diverse samples for training.

Table 7
The complexity of the architecture used in every experiment is expressed in terms of the number of parameters (in Million) and GFlops. EF and MF stand for early fusion and middle fusion respectively. In the fusion techniques, R, H, and D stand for RGB, HS↑, and DTM respectively, representing the combination of modalities adopted in the corresponding experiment.

Experiments	Parameters (Million)	GFLOPs
RGB	14.3	76.43
HS↑	14.9	202.89
DTM	14.3	75.02
EF RH	14.9	205.01
EF RD	14.3	77.14
EF RHD	14.9	205.71
MF RH	15.0	164.74
MF RD	14.8	164.74
MF RHD	15.3	209.95

The proposed dataset presents five modalities: RGB, panchromatic, VNIR, SWIR, and DTM and two labelings: the Land Cover with eight classes and the Soil Agricultural Use with 10 classes. To the best of our knowledge, this dataset is the biggest and most diverse dataset for RS semantic segmentation as it includes a high cardinality of images for all the modalities. Specifically, our dataset provides 1502 tiles and an extension of around 1332 km^2. The characteristics of this dataset, both in terms of spatial resolution and number of spectral bands, allow an important step for future studies, thus enabling the scientific community to explore the use of multi-modality in remote sensing.

Furthermore, we have investigated the advantages of these modalities. On the first hand, the scope of this analysis was to understand if the combination of complementary modalities can outperform the use of a single RGB modality, and, on the second hand, to provide a baseline for multi-modal RS semantic segmentation on the proposed dataset.

Summarizing, the main findings of the experimental investigation are:

- the empirical proof that the fusion of multiple modalities improves semantic segmentation accuracy compared to using a single-modality in both Land Cover and Soil Agriculture Use;
- the demonstration through empirical evidence that employing a middle fusion strategy enhances the effectiveness of multi-modality;
- the empirical evidence of the effectiveness of hyperspectral data in Soil Agricultural Use;
- the evaluated multi-modal deep networks require a number of parameters that is almost the same as single-modality deep networks.

Plenty of challenges connected to semantic segmentation are still open, and we think this dataset can become the first step in the right direction. This dataset can also help investigate open issues such as hyperspectral pansharpening, dimensionality reduction of high cardinality data, and spatio-temporal fusion of modalities.

One of the challenges refers to the refinement of the semantic labeling. The effect of noisy labels on model performance is partially mitigated by the intrinsic generalization capabilities of deep neural networks that are able to learn anomalous patterns and discard them. However, in future work, we plan to further refine the labeling, reducing the noisy labels and balancing the low-represented classes by using the dataset itself for a semi-supervised labelization of the background. This can be obtained by initially train a model with a small set of labeled data and then using such a model to predict labels for unlabeled areas, creating pseudo-labels. Combining these pseudo-labels with the original labeled dataset and retraining the model in an iterative manner allows the accuracy of the model to be improved over time.

Another challenge regards the generalization of the deep neural models. Our dataset, as for other remote sensing datasets, lacks of generalization ability, since it has been collected in a specific region and it has been acquired in almost clear sky condition. Thus, a deep network model that is trained on the proposed dataset may work properly in geographical areas only if those are very similar, in terms of geomorphology and sky conditions, to the one under study. To demonstrate the generalization of the proposed models on other geographic locations we should collect more data. However, the collection of a new set of data like the one we proposed is a very time-demanding activity, because it requires also time for pre-processing and labeling. In our future work, we intend to increase the extension of the dataset and its variability with the depth and attention it deserves.

In addition, another relevant conclusion from the evaluations, that points to one of the future directions in which research should move, is the need to study and develop new strategies for merging different information. The results show that using diverse strategies leads to a significant discrepancy in how the sources are treated and what both overall and class-focus performance is achieved. As a direct consequence, investigating more sophisticated fusion strategies and their repercussion on the segmentation is of fundamental relevance. In particular, we think that a good starting point for further experimentation would be to focus on the specific sources and classes. In our view, emphasis should be placed on designing fusion models that integrate each modality, while also considering the distinctive characteristics of each semantic class. We plan to use this data to continue investigating the field of RS semantic segmentation and to further exploit the usefulness of the HS and the DTM. This can be achieved only by studying and understanding more in-depth how to handle each modality to extract the best possible and appropriate information. This remains a future challenge that reflects the complexity of combining different sources that present different proprieties (e.g. time stamp, resolution, etc.), remarking once again the existence of multi-modal dataset as a desideratum.

CRediT authorship contribution statement

Mirko Paolo Barbato: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Flavio Piccoli:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Paolo Napoletano:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Research partially developed in the context of the project PIGNOLETTO – Call HUB Ricerca e Innovazione CUP (Unique Project Code) n. E41B20000050007, co-funded by POR FESR 2014–2020 (Programma Operativo Regionale, Fondo Europeo di Sviluppo Regionale – Regional Operational Programme, European Regional Development Fund).

Research partially supported by the MUR under the grant “Dipartimenti di Eccellenza 2023–2027” of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

The authors are grateful to Prof. Begüm Demir for the valuable comments and stimulating discussions.

Code availability

The code for the dataset pre-processing and for running the experiments, for both training and testing of the model, can be found at https://github.com/mpBarbato/Ticino-RS-Dataset.

Appendix A. Additional information on Ticino dataset

In this appendix, we will describe additional information about the Ticino dataset.

A.1. RGB and PRISMA disalignment

The alignment of the RGB and PRISMA sources has been done with an interactive approach that involved the selection of more than 700 correspondent pairs of Ground Control Points between the RGB and PAN images, and the following estimation of a Thin Plate Spline Transformation for the geometric correction. The selection and the transformation were applied using QGIS Desktop software (qgi, 2023). Fig. A.9(a) and (b) show the result of the alignment procedure. In the figures, two crops, considering RGB and panchromatic modalities, are overlapped to show the difference between before (A.9(a)) and after (A.9(b)) the alignment.

A.2. Training, validation and test splits

In this subsection, we discuss the class statistics for each labeling by including the division into training, validation and test. Fig. A.10 offers a deeper analysis of the Land Cover labeling. The first column shows the number of pixels belonging to each class, while the second column the number of pixels per label for all the three sets in which the dataset has been divided dataset, namely the training (a), the validation (b), and the test (c) set.

Finally, as before, a deeper analysis of the SAU distribution is proposed in Fig. A.11. The first column represents the number of pixels belonging to each class, while the second column the number of pixels per label for all three sets: training (a), the validation (b), and the test (c) set.

A.3. Refinement of the original labeling

In this subsection, we will describe the refinement process of the original labelings to achieve the two final ground truths for semantic segmentation.

The final dataset has been collected by merging information from Open Street Map (ope, 2023) and the Italian Agenzie delle Entrate (age, 2023), augmenting them with the creation of the *Water* labeling. As described in Section 3.1, the dataset consists of 8 classes: *Background*, *Building*, *Road*, *Residential*, *Industrial*, *Forest*, *Farmland*, and *Water*.

Background, *Residential*, *Park*, *Industrial*, and *Forest* originally derived from the OSM labeling (ope, 2023). The original OSM segmentation includes 22 classes: *Background*, *Buildings*, *Forest*, *Residential*, *Farmland*, *Parking*, *Industrial*, *Stadium*, *Meadow*, *Pond*, *Park*, *Square*, *Harbour*, *Airport*, *Bridge*, *Beach*, *Industrial harbour*, *Baseball*, *Desert*, *Rock*, *Glacier*, and *River*. After having divided the area under investigation into 1808 tiles, we have decided to discard the classes with low representations in terms of the number of image samples: *Harbour*, *Airport*, *Bridge*, *Beach*, *Industrial harbour*, *Baseball*, *Desert*, *Rock*, *Glacier*, and *River*. As a consequence, 306 samples have been discarded because they mainly included the *Background* class.

Building and *Road* labelings have been collected by the Italian Agenzie delle Entrate (age, 2023). The former has been inserted in the dataset as a substitute for the *Building* labeling of OSM because it is more accurate and complete in the area considered, while the latter was not present in the original OSM labeling.

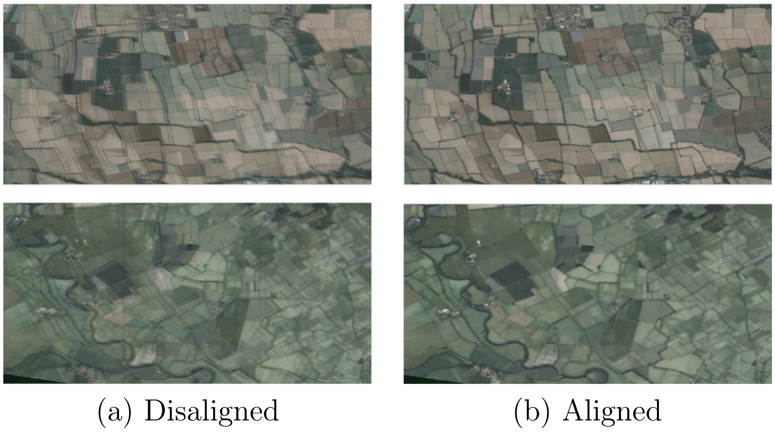
Finally, *Water* is a combination of the *Pond* segmentation provided by OSM and a manual labeling provided by the authors of the Ticino River.

The original Soil Agricultural Use labeling has been acquired from the Geoportal of Lombardia region (geo, 2023) and consisted of the following 22 classes: *Background*, *Other agricultural crops*, *Other cereals*, *Beet*, *Forests and tree crops*, *Nursery crops*, *Horticultural crops*, *Forage crops*, *Fruit crops*, *Corn*, *Olive tree*, *Industrial plants*, *Rice*, *Seeds*, *Tainted and uncultivated*, *Fallow land*, *Vine*, *Man-made areas*, *Natural barren areas*, *Water bodies*, *Unclassifiable agricultural land*, and *Natural vegetation*.

Other cereals, *Floriculture crops*, *Horticultural crops*, *Fruit crops*, *Vine*, *Beet*, and *Olive-tree* labels have been removed due to the low representation in the area considered. While *Forest and tree crops* and *Natural barren areas* have been respectively joined with the *Natural vegetation* and *Water bodies* as they have a similar semantic meaning. Finally, *Unclassifiable agricultural land*, *Tares and uncultivated*, and *Fallow land* were merged with the *Background* class because the semantic meaning was not clearly defined. The final labeling resulting from the cleaning process includes 10 classes as follows: *Background*, *Other agricultural crops*, *Forage crops*, *Corn*, *Industrial plants*, *Rice*, *Seeds*, *Man-made areas*, *Water bodies*, and *Natural vegetation*.

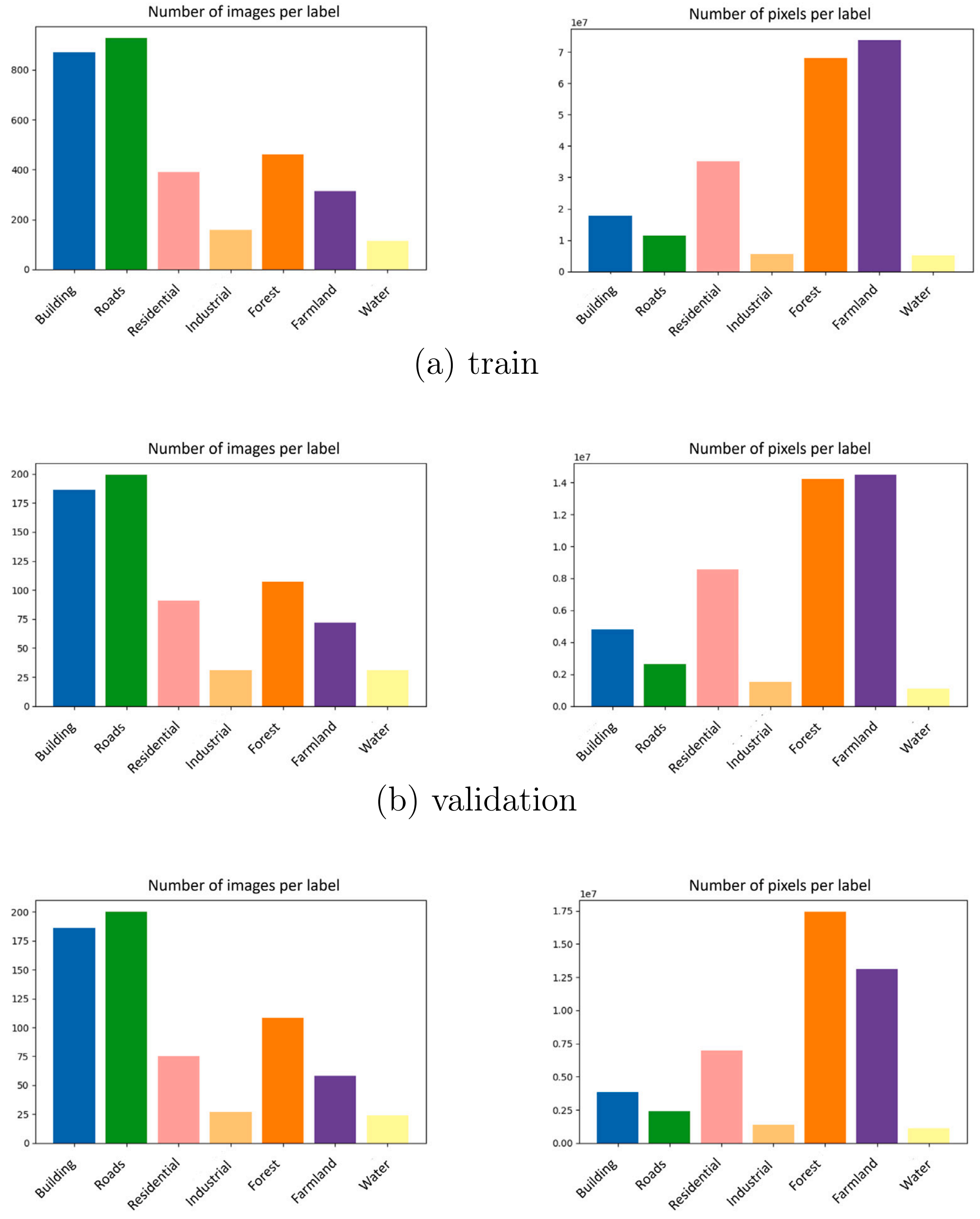
Appendix B. U-shaped network with ResNet18 backbone

In this section, we report a summary of the U-shaped neural network (with ResNet18 as the backbone) and its main components when an RGB image is considered as input. In Table B.8 is reported the general architecture, in Table B.9 the general structure of a BasicBlock and in Table B.10 the general structure of a DecoderBlock.



(a) Disaligned (b) Aligned

Fig. A.9. Disalignment of RGB, Panchromatic, and hyperspectral data. The figure shows two RGB and panchromatic crops overlapped before (left) and after (right) the alignment operations.

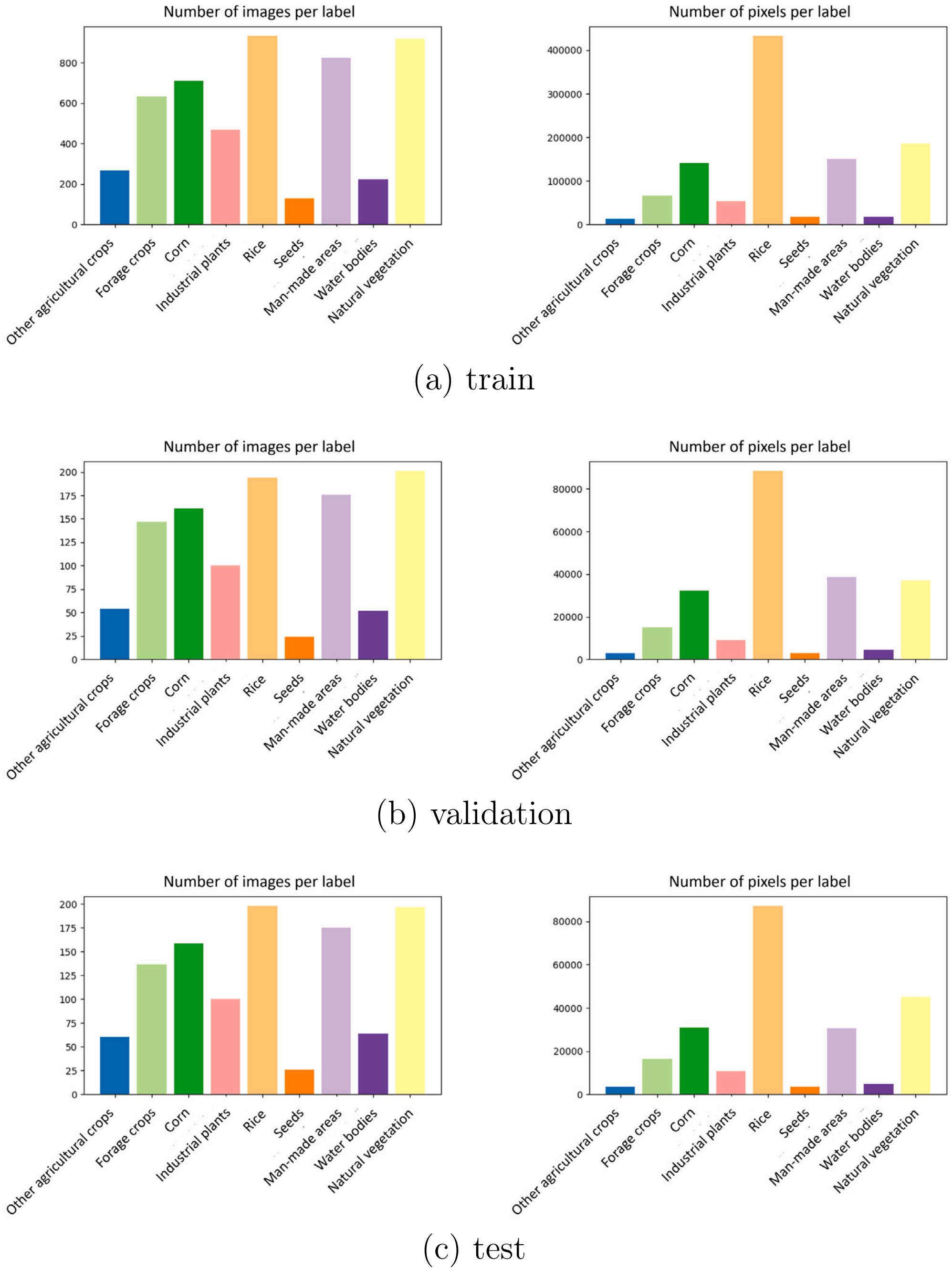


(a) train

(b) validation

(c) test

Fig. A.10. Distribution of the Land Cover split of the dataset in training (first row), validation (second row), and test (third row) sets. The first column represents the number of images per class (without Background). The second column represents the number of pixels per class (without Background).



(a) train

(b) validation

(c) test

Fig. A.11. Distribution of the Soil Agricultural Use split of the dataset in training (first row), validation (second row), and test (third row) sets. The first column represents the number of images per class (without Background). The second column represents the number of pixels per class (without Background).

Table B.8
U-shaped architecture with ResNet18 backbone.

Layer (depth-idx)	Output shape
ResNetEncoder: 1–1	[−1, 3, 256, 256]
Conv2d: 2–1	[−1, 64, 128, 128]
BatchNorm2d: 2–2	[−1, 64, 128, 128]
ReLU: 2–3	[−1, 64, 128, 128]
MaxPool2d: 2–4	[−1, 64, 64, 64]
Sequential: 2–5	[−1, 64, 64, 64]
BasicBlock: 3–1	[−1, 64, 64, 64]
BasicBlock: 3–2	[−1, 64, 64, 64]
Sequential: 2–6	[−1, 128, 32, 32]
BasicBlock: 3–3	[−1, 128, 32, 32]
BasicBlock: 3–4	[−1, 128, 32, 32]
Sequential: 2–7	[−1, 256, 16, 16]
BasicBlock: 3–5	[−1, 256, 16, 16]
BasicBlock: 3–6	[−1, 256, 16, 16]
Sequential: 2–8	[−1, 512, 8, 8]
BasicBlock: 3–7	[−1, 512, 8, 8]
BasicBlock: 3–8	[−1, 512, 8, 8]
UnetDecoder: 1–2	[−1, 16, 256, 256]
Identity: 2–9	[−1, 512, 8, 8]
ModuleList: 2	[]
DecoderBlock: 3–9	[−1, 256, 16, 16]
DecoderBlock: 3–10	[−1, 128, 32, 32]
DecoderBlock: 3–11	[−1, 64, 64, 64]
DecoderBlock: 3–12	[−1, 32, 128, 128]
DecoderBlock: 3–13	[−1, 16, 256, 256]
SegmentationHead: 1–3	[−1, 18, 256, 256]
Conv2d: 2–10	[−1, 18, 256, 256]
Identity: 2–11	[−1, 18, 256, 256]
Activation: 2–12	[−1, 18, 256, 256]
Identity: 3–14	[−1, 18, 256, 256]

Table B.9
BasicBlock general structure. N represents the number of channels, the height, and the width of the input.

Layer	Output shape
BasicBlock	[−1, N, N, N]
Conv2d	[−1, N, N, N]
BatchNorm2d	[−1, N, N, N]
ReLU	[−1, N, N, N]
Conv2d	[−1, N, N, N]
BatchNorm2d	[−1, N, N, N]
ReLU	[−1, N, N, N]

Table B.10
DecoderBlock general structure. C represents the number of channels of the input, while N represents the height and the width of the input.

Layer	Output shape
DecoderBlock	[−1, C, N, N]
Conv2d	[−1, C, N, N]
BatchNorm2d	[−1, C, N, N]
ReLU	[−1, C, N, N]

References

age (2023). Consultazione cartografia catastale, 2015. https://www.agenziaentrate.gov.it/portale/web/guest/schede/fabbricatiterreni/consultazione-cartografia-catastale/servizio-consultazione-cartografia.

Aiazzi, B., Baronti, S., & Selva, M. (2007). Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, *45*, 3230–3239.

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., et al. (2023). Transformers in remote sensing: A survey. *Remote Sensing*, *15*, 1860.

Arora, S. K. (2018). Spacenet information. Medium.

Barbato, M. P., Napoletano, P., Piccoli, F., & Schettini, R. (2022). Unsupervised segmentation of hyperspectral remote sensing images with superpixels. *Remote Sensing Applications: Society and Environment*, *28*, Article 100823.

Baumgardner, M. F., Biehl, L. L., & Landgrebe, D. A. (2015). 220 Band aviris hyperspectral image data set: June 12 1992indian pine test site 3. *Purdue University Research Repository*, *10*, 991.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*, 2–16.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: fast and flexible image augmentations. *Information*, *11*, 125.

Dechesne, C., Mallet, C., Le Bris, A., & Gouet-Brunet, V. (2017). Semantic segmentation of forest stands of pure species as a global optimization problem. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*, 141.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., et al. (2018). Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 172–181).

dos Santos, J. A., Penatti, O. A., Gosselin, P.-H., Falcão, A. X., Philipp-Foliguet, S., & Torres, R. d. S. (2014). Efficient and effective hierarchical feature propagation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*, 4632–4643.

Dst (2023). Dstl satellite imagery feature detection. https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection.

Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, *32*, 829–864.

Gavrilyuk, K., Sanford, R., Javan, M., & Snoek, C. G. (2020). Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 839–848).

geo (2023). Geoportal of lombardy region. https://www.geoportale.regione.lombardia.it.

Hasan, M. K., Lee, S., Rahman, W., Zadeh, A., Mihalcea, R., Morency, L.-P., et al. (2021). Humor knowledge enriched transformer for understanding multimodal humor. *Vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12972–12980).

He, Q., Sun, X., Diao, W., Yan, Z., Yao, F., & Fu, K. (2023). Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling. *IEEE Transactions on Image Processing*, *32*, 1474–1487. http://dx.doi.org/10.1109/TIP.2023.3245324.

Isp (2023). ISPRS 2D semantic labeling, 2018. https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx#!.

Jadhav, J. K., & Singh, R. (2018). Automatic semantic segmentation and classification of remote sensing data for agriculture. *Mathematical Models in Engineering*, *4*, 112–137.

Jiang, X., Ma, J., Xiao, G., Shao, Z., & Guo, X. (2021). A review of multimodal image matching: Methods and applications. *Information Fusion*, *73*, 22–71.

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, *14*, 778–782.

Lateef, F., & Ruichek, Y. (2019). Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, *338*, 321–348.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., et al. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, *112*, Article 102926.

Li, Y., Zhang, J., Cheng, Y., Huang, K., & Tan, T. (2017). Semantics-guided multi-level rgb-d feature fusion for indoor semantic segmentation. In *2017 IEEE international conference on image processing* (pp. 1262–1266). IEEE.

Lin, J., Yang, A., Zhang, Y., Liu, J., Zhou, J., & Yang, H. (2020). Interbert: Vision-and-language interaction for multi-modal pretraining. arXiv preprint arXiv:2003.13198.

Loncan, L., De Almeida, L. B., Bioucas-Dias, J. M., Briottet, X., Chanussot, J., Dobigeon, N., et al. (2015). Hyperspectral pansharpening: A review. *IEEE Geoscience and Remote Sensing Magazine*, *3*, 27–46.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32.

M Graña, B. A., & Veganzons, M. A. (2020). Hyperspectral remote sensing scenes. Ehu.

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE international geoscience and remote sensing symposium* (pp. 3226–3229). IEEE.

mic (2023). *Microsoft bing maps*. MBM, https://www.bing.com/maps/.

Mohanty, S. P., Czakon, J., Kaczmarek, K. A., Pyskir, A., Tarasiewicz, P., Kunwar, S., et al. (2020). Deep learning for understanding satellite imagery: An experimental survey. *Frontiers in Artificial Intelligence*, 3.

ope (2023). OpenStreetMap. https://www.openstreetmap.org.

Palhamkhani, F., Alipour, M., Dehnad, A., Abbasi, K., Razzaghi, P., & Ghasemi, J. B. (2023). Deepcompoundnet: enhancing compound–protein interaction prediction with multimodal convolutional neural networks. *Journal of Biomolecular Structure and Dynamics*, 1–10.

Parida, K. K., Srivastava, S., & Sharma, G. (2022). Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3347–3356).

pri (2023). Prisma satellites data. https://www.asi.it/scienze-della-terra/prisma/.

qgi (2023). QGIS. https://www.qgis.org/it/site/.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., et al. (2012). The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012) Nr. 1*, *1*, 293–298.

Santiago, J. G., Schenkel, F., Gross, W., & Middelmann, W. (2020). An unsupervised labeling approach for hyperspectral image classification. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 407–415.

SIA (2023). Carta dell'uso agricolo dati. https://www.cartografia.regione.lombardia.it/metadata/raster/lyr/UTM/ags/doc/Utilizzo_agroforestale_da_dati_SIARL.pdf.

Van Etten, A., Lindenbaum, D., & Bacastow, T. M. (2018). Spacenet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232.

Vivone, G., Garzelli, A., Xu, Y., Liao, W., & Chanussot, J. (2022). Panchromatic and hyperspectral image fusion: Outcome of the 2022 whispers hyperspectral pansharpening challenge. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 166–179.

Volpi, M., & Ferrari, V. (2015). Semantic segmentation of urban scenes by learning local class interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1–9).

Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., et al. (2016). Torontocity: Seeing the world with a million eyes. arXiv preprint arXiv:1612.00423.

Xie, T., Wang, K., Lu, S., Zhang, Y., Dai, K., Li, X., et al. (2023). Co-net: Learning multiple point cloud tasks at once with a cohesive network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3523–3533).

Xie, T., Wang, S., Wang, K., Yang, L., Jiang, Z., Zhang, X., et al. (2023). Poly-pc: A polyhedral network for multiple point cloud tasks at once. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1233–1243).

Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuan, X., Shi, J., & Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, *169*, Article 114417.

Zhan, X., Wu, Y., Dong, X., Wei, Y., Lu, M., Zhang, Y., et al. (2021). Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11782–11791).

Zini, S., Barbato, M. P., Piccoli, F., & Napoletano, P. (2023). Deep learning hyperspectral pansharpening on large scale prisma dataset. arXiv:2307.11666.